Evaluation of missing data techniques for in-car automatic speech recognition

Y. Wang[1], R. Vuerinckx[2], J. Gemmeke[3], B. Cranen[3], H. Van hamme[1]

[1] *ESAT Department, Katholieke Universiteit Leuven, Belgium,*
Email: yujun.wang@esat.kuleuven.be, hugo.vanhamme@esat.kuleuven.be

[2] *Nuance Communications, Merelbeke, Belgium, Email: rudi.vuerinckx@nuance.com*

[3] *Radboud Universiteit Nijmegen, The Netherlands, Email: j.gemmeke@ let.ru.nl, b.cranen@let.ru.nl*

Introduction

One of the major concerns in deploying speech recognition applications is the lack of robustness of the technology. One key aspect is the sensitivity to stationary or non-stationary background noise. Many approaches to noise robust speech recognition have been proposed before. Some modify the front-end signal processing of the recogniser while others work on the back-end, i.e. modelling and decoding. Stationary noise can be handled with techniques such as spectral subtraction while non-stationary noise is difficult to remove. Approaches that can handle non-stationary noise such as parallel model combination [1], blind source separation [2] and model-based decomposition [3], require explicitly modelling the statistics of the noise.

Besides these methods, multi-style training is well known to give superior recognition performance. The acoustic model is trained on corrupted speech. It is a very effective and practical way to improve the noise robustness of a speech recogniser hence is widely applied on many commercial products such as VOCON 3200 of Nuance communication. Additionally, multi-style training can easily be combined with speech enhancement schemes. In [11], the author gives a survey of multi-style training. Here "style" especially refers to the acoustic environmental condition, while "multi" means that the training data consists of speech embedded in different types of noise,. The noisy training data can either be obtained by direct recording or by artificially adding particular kinds of noise into the training data. The former way is expensive for car application because the data has to be collected in a real driving condition while the latter is more economical but less effective. Finally, multi-style training blurs the acoustic models and makes them less discriminative, which could lead to loss in accuracy.

Missing Data Techniques (MDT) based speech recogniser only relies on the knowledge of the clean speech. It increases the noise robustness of the recogniser by curing the mismatch between the acoustic model and the observations when stationary or non-stationary noise is present. Hence training various acoustic models according to the environment noise in the target application is not a must in MDT and the cost is greatly reduced with respect to the multi-style training.

MDT relies on the property that some regions in the spectrogram are dominated by the speech signal to be recognized, while other regions are corrupted by unexpected noise. MDT solutions tell, in the frontend, which frequency components of a frame of speech spectrum are corrupted by noise (unreliable) and, in the backend, how to impute them with the acoustic model and non-corrupted components (reliable). This reliability information is represented in a spectral mask which is estimated from the noisy data. In the backend, the imputation step replaces the unreliable observed speech by clean speech estimation to assure a good match between the clean acoustic model and the corrupted data.

In [4], MDT is applied based on continuous density Hidden Markov Model (HMM). However, the acoustic model must be expressed as a mixture of Gaussians with diagonal covariance matrix in the log spectral domain, where the masks are represented. MDT recognisers working with spectral features (SMDT) exhibit a loss of accuracy while using the cepstral domain increases accuracy due to the property that the Discrete Cosine Transform (DCT) decorrelates the log-spectra. In [5], the cepstral MDT (CMDT) system shows superior noise robustness in comparison with a SMDT system. In CMDT, the cosine transformation is applied on the diagonal cepstral covariance matrix. Imputing clean speech based on observed noisy speech and a Gaussian mixture of a clean acoustic model with the assumption that the difference between noisy speech and the clean speech to be imputed is non-negative, requires solving a Non-Negative Least Square (NNLSQ) problem. Solving a NNLSQ leads to a severe computational load. In [6], an alternative MDT formulation through the introduction of the PROSPECT features is presented. It reduces the computational requirements of NNLSQ and maintains the accuracy of CMDT at the same time.

In the next section, MDT masks and imputation are first reviewed, followed by the introduction of the PROSPECT features. The resulting implementation is benchmarked on the SpeechDat CAR Flemish database [10] and compared with the results from the VOCON 3200 (version 2.6) to see how far the MDT is away from a state-of-the-art speech recogniser.

Missing Data Techniques

In MDT, when the speech signal is contaminated by additive noise, a spectral mask indicates at each time frame which spectral components are labelled as missing or unreliable (dominated by noise) and which are reliable (dominated by speech). A detail survey of MDT masks can be found in [8]. Using a mask, a D-dimensional vector $\mathbf{y}$ containing the spectral observations of frame t can be split into an unreliable part $\mathbf{y}_u$ and a reliable part $\mathbf{y}_r$:

$$\mathbf{y}' = [\mathbf{y}_u' \quad \mathbf{y}_r'] \tag{1}$$

The reliable components of the clean speech $\mathbf{s}$ are estimated as the noisy observation $\mathbf{y}$. The assumption that the noise is

additive yields the constraint in each unreliable filter bank channel that the clean speech to be imputed $\mathbf{s}_u$ is bounded:

$$\mathbf{s}_u \leq \mathbf{y}_u \quad (2)$$

In data imputation, the unreliable components $\mathbf{s}_u$ are estimated under constraint (2), producing a complete observation vector $\mathbf{s}$. The estimation uses a (Gaussian) model of speech, which is given by the current search hypothesis, i.e. the imputation is integrated with the decoding process. In [4], the authors suggest a diagonal covariance MDT solution in the log spectral domain where the mask is presented. The bounded imputation carried out per Gaussian and per frame is straightforward in this case. For the sake of better accuracy, cepstral MDT is introduced in [5], where each Gaussian is evaluated separately as well. Hence the acoustic model can be evaluated on complete spectral data by mixture-wise imputation. The minimization of the cost function of a mixture

$$(\mathbf{s}-\boldsymbol{\mu}_s)'\mathbf{C}'\Sigma_c^{-1}\mathbf{C}(\mathbf{s}-\boldsymbol{\mu}_s) = (\mathbf{y}-\mathbf{x}-\boldsymbol{\mu}_s)'\mathbf{C}'\Sigma_c^{-1}\mathbf{C}(\mathbf{y}-\mathbf{x}-\boldsymbol{\mu}_s) \quad (3)$$

becomes an NNLSQ problem over $\mathbf{x}$, which is the difference between the imputed speech and the noisy observation. $\mathbf{C}$ is the DCT matrix and $\boldsymbol{\mu}_s$ denotes the spectral mean of the mixture. The NNLSQ problem can be solved by gradient descent. However the computational load is significantly higher than for SMDT because we have to calculate the inverse covariance or the precision matrix $\mathbf{C}'\Sigma^{-1}\mathbf{C}$ during decoding. To save computational efforts, PROSPECT features are introduced.

PROSPECT Features

The PROSPECT [6] feature is composed of a cepstral vector *c* and a projected vector *d* formulated by:

$$\mathbf{p} = \begin{bmatrix}\mathbf{C}_K \\ \mathbf{P}_\perp\end{bmatrix}\mathbf{s} = \begin{bmatrix}\mathbf{c} \\ \mathbf{d}\end{bmatrix} \quad (4)$$

$\mathbf{s}$ is the log spectrum. $\mathbf{C}_K$ is the K by D orthonormal DCT matrix. K is much lower than 13, e.g. K=4, which means only the lower order spectral correlations are captured by the K cepstral coefficients statistically. $\mathbf{P}_\perp=\mathbf{I}-\mathbf{C}_K'\mathbf{C}_K$ hence $\mathbf{d}=\mathbf{s}-\mathbf{C}_K'\mathbf{C}_K\mathbf{s}$, which is spectral residual after $\mathbf{c}$ is removed from $\mathbf{s}$.

The cost function to be minimised is:

$$(\mathbf{s}-\boldsymbol{\mu}_s)'\left[\mathbf{C}_K'\Sigma_c^{-1}\mathbf{C}_K + \alpha\mathbf{P}_\perp'\Sigma_d^{-1}\mathbf{P}_\perp\right](\mathbf{s}-\boldsymbol{\mu}_s) = (\mathbf{s}-\boldsymbol{\mu}_s)'\mathbf{A}(\mathbf{s}-\boldsymbol{\mu}_s) \quad (5)$$

α is the stream weight of d and is 0.5 in our case. $\mathbf{A}$ is the precision matrix. When using gradient decent to solve NNLSQ, the gradient vector is $\mathbf{A}(\mathbf{s}-\boldsymbol{\mu}_s)$ and can be calculated using the lower order matrices $\mathbf{C}_K$, $\mathbf{C}_K'$, diagonal matrices Σ_c, Σ_d and the vector $\mathbf{s}-\boldsymbol{\mu}_s$. Hence computational efforts are saved by avoiding a series of multiplications with the full precision matrix as is required in the imputations for CMDT. The PROSPECT based recogniser has been proven to have an equivalent performance as the CMDT recogniser.

Experiments

The PROSPECT MDT recogniser and the PROSPECT recogniser with MDT disabled, as well as the VOCON 3200 recogniser are tested in the experiments on the SpeechDat CAR Flemish database.

Recognition tasks

The database consists of noisy utterances recorded in different driving conditions. Speech is recorded using four channels from close, medium and far field microphones. The test utterances are binned per 5 dB SNR in the range of 0~30dB SNR. Four grammars are included in the experiments, isolated words, spelling, natural number and "when" grammars. The isolated words grammar has 637 words or commands in its dictionary. The spelling grammar allows an arbitrary number of repetitions of a letter of the alphabet. The natural number grammar allows the speaker to say a number in the range from 1 to 999,999 in a natural way as well as money amounts in the same range. The "when" grammar allows the speaker to specify a point in time. This can be a date, or a time, but also relative indications like "tomorrow" or "in one week" are allowed.

Recognisers

The PROSPECT MDT recogniser, the PROSPECT recogniser without MDT and the VOCON 3200 engine from Nuance Communications are tested. The Flemish PROSPECT acoustic model set is trained from 60 hours of clean data only. Channel normalisation [9] and the VQ mask [7] are used for the MDT recogniser. The VOCON 3200 ASR engine is a small-footprint engine, using MFCC based features and HMM models. It contains speech enhancement techniques to cope with stationary or slowly varying background noise. Its training data includes in-car recorded samples, i.e. it uses the multi-style training approach in tandem with noise reduction techniques. Other differences exist between the two systems: some vocabulary items of the VOCON recogniser use whole word HMM models, while the MDT system uses triphones throughout.

Results

Figure 1-4 shows the experiments results per 5dB SNR bin (regardless of channel) of the above four grammars.

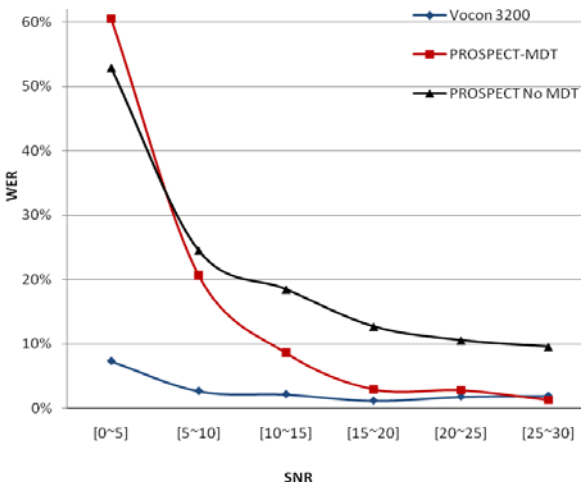

Figure 1: WER of the isolated words obtained with VOCON 3200 recogniser, the PROSPECT MDT based recogniser and the PROSPECT recogniser without MDT.

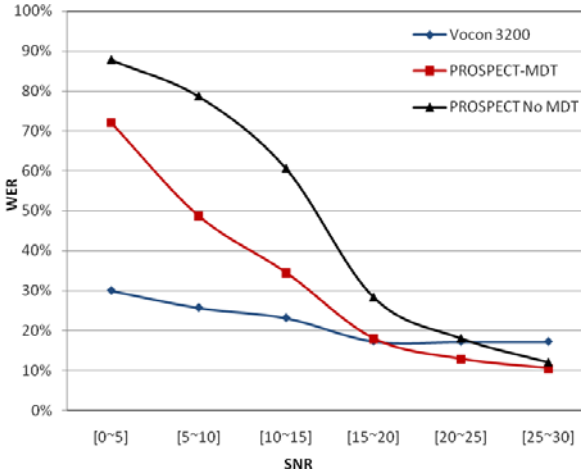

Figure 2: WER of the spelling grammar obtained with VOCON 3200 recogniser, the PROSPECT MDT based recogniser and the PROSPECT recogniser without MDT.

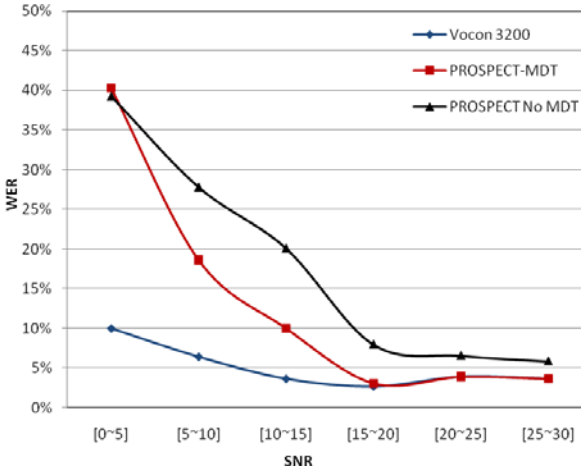

Figure 3: WER of the natural number grammar obtained with VOCON 3200 recogniser, the PROSPECT MDT based recogniser and the PROSPECT recogniser without MDT.

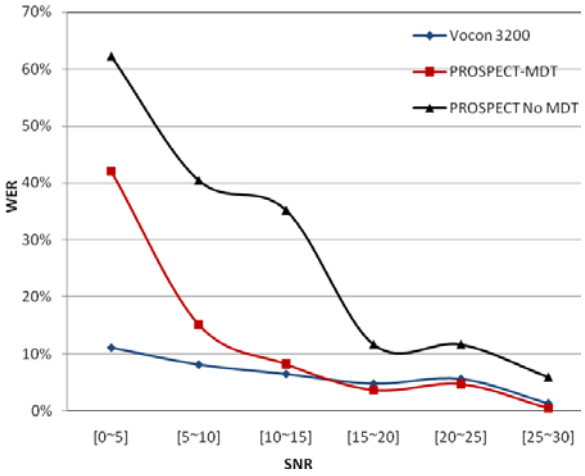

Figure 4: WER of the "when" grammar obtained with VOCON 3200 recogniser, the PROSPECT MDT based recogniser and the PROSPECT recogniser without MDT.

Discussion and future work

In almost all cases, the MDT recognizer outperforms its variant that is not using noise compensation, even at high SNR. At an SNR above 15dB, the MDT system and the VOCON3200 recognizer have comparable performance. At an SNR below 15 dB, the MDT system is outperformed by the VOCON3200. This result is not surprising given the different assumptions about the noise, the differences in feature representations, the differences in amount and type of training data, etc. The errors in the estimated masks are also one of the potential reasons.

Despite the performance gap between the MDT recogniser and the VOCON recogniser, MDT system is versatile and a lot cheaper in data collection and is not tuned to its operating environment: it does not require collecting in-car training data from the acoustic environment that it will be deployed in.

Since we conclude from this result analysis that multi-style training works well when dealing with corrupted speech, using multi-style trained acoustic model in MDT is a potential further direction, though this compromises on the foundation for MDT that we want to make minimal assumptions about the noise. Moreover, the accuracy of mask estimation should be increased. Hereto, longer context information will also be considered.

Though we did not provide any analysis of this aspect, MDT-based systems are also computationally more expensive.

Acknowledgements

This research is financed by the MIDAS project of the Nederlandse Taalunie under the STEVIN programme.

References

[1] Gales, M., "Model-based techniques for noise robust speech recognition", PhD thesis, Univ. Of Cambridge, Sept, 1995.

[2] Ikeda, S. and Murata, N., "An approach to blind source separation of speech signals", Proc. International Conference on Artificial Neural Networks, vol. 2. pp. 761-766, 1998.

[3] Stouten, V., Van hamme, H., Demuynck, K., Wambacq, P., "Robust speech recognition using model-based feature enhancement", Proc. Eurospeech, Geneva, pp. 17-20, Sept. 2003.

[4] Cooke, M., Green, Ph., Josifovski, L., Vizinho, A. "Robust automatic speech recognition with missing and unreliable acoustic data", Speech Communication 34 (2001), pp. 267-285

[5] Van hamme, H., "Robust Speech Recognition Using Missing Feature Theory in the Cepstral or LDA Domain," Proc. Eurospeech, Geneva, pp. 3089-3092. 2003.

[6] Van hamme, H., “PROSPECT Features and their Application to Missing Data Techniques for Robust Speech Recognition”, Proc. International Conference on Spoken Language Processing, volume I, 101-104, 2004.

[7] Van Segbroeck, M., Van hamme, H., “Vector-Quantization Based Mask Estimation for Missing Data Automatic Speech Recognition”, Proc. Interspeech, 910-913, 2007.

[8] Cerisara, C., Demange, S., Haton, J.-P., “On noise masking for automatic missing data speech recognition: a survey and discussion”, Computer Speech and Language, 21(3):443-457, July 2007.

[9] Van Segbroeck, M., Van hamme, H., “Handling Convolutional Noise in Missing Data Automatic Speech Recognition”, Proc. International Conference on Spoken Language Processing, pages 2562-2565, Pittsburgh, U.S.A., Sept. 2006.

[10] Moreno, A., Lindberg, B., Draxler, C., Richard, G., Choukri, K., Allen, J., Euler. S., “Speechdat-car: A large speech database for automotive environments”. LREC, 2000.

[11] Huang X., Hon, H.W., Reddy, R., “Spoken Language Processing: A Guide to Theory, Algorithm, and System Development”, Prentice Hall, NJ, 2001